From One-Hot to Semantic Encoding: Entity Embedding for Small and Heterogeneous Digital Humanities Datasets

Isabelle Gribomont
CENTAL/UCLouvain
Place de l'Université 16
1348 Louvain-la-Neuve, Belgique
isabelle.gribomont@uclouvain.be

Abstract

This paper investigates the use of semantic encoding for the analysis of heterogeneous digital literature metadata. Drawing on two databases of Latin American digital literature, *Archivo de Literatura Digital en América Latina* and the *Atlas da Literatura Digital Brasileira*, we compare traditional one-hot encoding with a semantically enriched representation derived from feature-value descriptions embedded in a continuous vector space. In contrast to one-hot encoding, which treats categorical values as orthogonal, semantic encoding models accounts for similarity between features, thereby mitigating vocabulary mismatch across databases. We evaluate both approaches using between-group centroid distances, and normalized centrality measures. Our results show that semantic encoding clarifies structural differentiation across genres and might smooth arbitrary differences introduced by differing vocabularies across databases. The findings suggest that semantic representations provide a more interpretable embedding space for small and taxonomically heterogeneous datasets. Beyond technical performance, the study suggests that embedding-based methods can support critical inquiry in digital humanities, enabling the examination of database bias, categorical patterns, and diachronic evolution within a unified semantic framework. Code is available at https://github.com/isag91/semantic-encoding-DH.

Keywords: Categorical Data, Entity Embeddings, Digital Humanities

1. Introduction

Digital humanities research increasingly relies on structured databases to document, classify, and analyze cultural production. In the field of digital literature, initiatives such as the *Archivo de Literatura Digital en América Latina - ADLAL* (Archive of Digital Literature in Latin America) and the *Atlas da Literatura Digital Brasileira - ALDB* (Atlas of Brazilian Digital Literature) provide curated metadata describing works according to genre, publication format, technical requirements, hardware dependencies, artistic techniques etc. (Botero Betancur, 2023; Athayde and ROCHA, 2022). These databases are invaluable for scholarly inquiry, yet they pose methodological challenges.

First, they are relatively small. Digital humanities datasets often contain hundreds rather than thousands of entries. Therefore, patterns of category co-occurrence are often insufficient to reliably infer semantic relationships between labels (e.g. the close conceptual relationship between *hypertext* and *hypermedia*). Second, they are heterogeneous: distinct databases use different vocabularies, taxonomies, and degrees of granularity. Third, many features are multi-valued (e.g., a work may require a computer and loudspeakers to be consulted).

These difficulties limits the ways in which these datasets can be meaningfully leveraged within the field. One-hot encoding treats categorical values as equidistant, ignoring semantic relationships between them. This limitation creates what has been termed a semantic gap in categorical representation (Yang et al., 2026). Likewise, when different databases refer to similar practices using distinct terminologies, their underlying semantic proximity is obscured.

Recent work has proposed leveraging generative large language models (LLMs) to enrich categorical representations with external semantic knowledge and/or transformer semantic encoder models to represent categorical descriptions. The potential of these approaches for humanities datasets—characterized by sparsity, conceptual overlap, and evolving vocabularies—remains underexplored.

This paper investigates whether semantic encoding of categorical metadata can improve geometric representation of data points in the semantic space in the context of two digital literature databases. To do so, we compare a traditional one-hot baseline with a semantic encoding approach in which feature-value labels are described using a large language model and embedded using a pre-trained sentence transformer.

Our central hypothesis is that semantic encoding offers methodological advantages for humanities datasets in at least three respects:

1. It captures conceptual proximity between labels that are treated as independent symbols in one-hot

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 147–152
11 May 2026.

representations.

2. It mitigates sparsity by importing external semantic knowledge.
3. It facilitates cross-database comparison despite divergent terminologies.

In the case of the two digital literature databases examined here, a semantically informed encoding could enable analyzes that would otherwise be distorted by heterogeneous vocabularies and the absence of explicit relationships between categorical values. Such an approach would make it possible to investigate whether certain databases occupy only specific subregions of the semantic space; whether works by authors of different genders or countries of origin are distributed differently across this space of practices; and whether diachronic patterns can be observed, such as the emergence of new regions over time.

2. Related Work

Traditional one-hot encoding represents categorical variables as orthogonal vectors, implicitly assuming that all categories are equally dissimilar. In machine learning, this limitation has motivated the development of learned embeddings for categorical variables, particularly in tabular data contexts (Guo and Berkhahn, 2016; Gorishniy et al., 2021). These methods learn semantic relationships from co-occurrence patterns within the dataset. Such inferences require a large amount of data to be reliable.

Large language models (LLMs) trained on large-scale corpora encode extensive distributional knowledge and have demonstrated strong performance across a wide range of semantic tasks (Brown et al., 2020; Devlin et al., 2019; Feng et al., 2024). Beyond text generation, such models can serve as structured semantic resources capable of producing definitions, contextual descriptions, and attribute-level interpretations. Recent work has highlighted the potential of LLMs as general-purpose knowledge interfaces (Bommasani et al., 2021; Petroni et al., 2019; AlKhamissi et al., 2022).

Building on these observations, recent approaches have proposed to leverage LLMs as external knowledge bases to semantically enrich datasets, especially in cases where intra-dataset co-occurrence signals are too sparse to infer meaningful relationships between categorical values. By querying LLMs at the feature–value level and generating structured descriptions or embeddings, it becomes possible to inject semantic information that is not recoverable from the data alone (Yang et al., 2026; Huesmann and Linsen, 2025; Hegselmann et al., 2023).

Within digital humanities, embedding-based methods have primarily been applied to textual corpora, supporting large-scale stylistic, thematic, or historical analysis (Underwood, 2019; Piper, 2018). However, comparatively less attention has been devoted to the modeling of metadata structures themselves. By leveraging semantically enriched embeddings at the metadata level, this study extends vector-space modeling beyond textual analysis and into the domain of structured descriptive information. This method has the potential to improve the investigation of database bias, categorical patterns, and diachronic shifts within a unified semantic framework, particularly in small and heterogeneous datasets where statistical and deep learning methods relying on within-dataset co-occurrence may be limited.

3. Methodology

3.1. Data Sources

We use two curated databases of digital literature. The Archivo de Literatura Digital en América Latina indexes works from Latin America, except Brazil, and contains 180 items. The Atlas da Literatura Digital Brasileira indexes works from Brazil and contains 149 items. Both databases describe digital literature using very similar metadata fields such as *genre*, *publication type*, *access hardware*, *technical requirements* etc. The databases differ slightly in taxonomy and naming conventions. To enable comparison, feature names were mapped to a shared schema in English, but differences in the vocabulary itself were maintained, besides translation.

Missing values were encoded using an explicit *no_information* label to preserve structural consistency rather than discarding incomplete entries.[1]

3.2. Semantic Description of Feature-Value Pairs

Following a similar approach to the ARISE framework (Yang et al., 2026), for each unique feature-value pair (e.g., *genre_poetry*, *publication_type_software*), we generated a structured description using ChatGPT-5.3.[2] Each description followed a fixed template: [CORE]: general definition of the value. [INDICATOR]: what this value indicates in the context of digital literature.[3]

[1] The data, LLM prompt and code are available at https://github.com/isag91/semantic-encoding-DH

[2] GPT-5.1 was found to offer the best performance compared to Claude Opus 4.5, DeepSeek V3.2 and Gemini 3 Pro by Yang et al. (2026).

[3] Examples: *access_hardware_computer*: [CORE] A computer is an electronic device that processes data and

A potential limitation of LLM-based semantic encoding lies in the stochastic nature of text generation, which may introduce variability in the resulting representations. To mitigate this issue, descriptions were generated at the feature-value level rather than per instance, ensuring that identical categorical values are consistently mapped to the same semantic representation. In addition, structured prompting was employed to constrain the form and content of the generated descriptions, reducing variability and emphasizing discriminative information. Finally, descriptions generated by a LLM could also be replaced by expert definitions. Once the descriptions are set, the semantic encoding pipeline is stable and deterministic.

3.3. Embedding Strategy

Once structured descriptions were generated for each unique feature-value pair, these texts were transformed into dense vector representations using the pre-trained sentence transformer *all-mpnet-base-v2* used in Yang et al. (2026). To convert variable-length text into fixed-dimensional vectors, a pooling strategy is adopted to weight tokens according to their activation levels. Following (Yang et al., 2026), we extract the last hidden state of the transformer, which provides a contextualized embedding for each token. A scalar activation score is computed for each token as the mean across embedding dimensions, and these scores are normalized with a softmax to produce attention weights. The embedding is computed as an attention-weighted sum.

When a work contained multiple values for the same feature, their corresponding embeddings were averaged, resulting in one vector per feature for each work. After computing feature-level embeddings, vectors corresponding to all features of a given work were concatenated to form a unified semantic representation. Concatenation was chosen over summation to preserve feature-specific structure and avoid conflating semantically distinct metadata dimensions. In this way, each feature occupies a stable subspace within the overall representation.

runs programs, typically with a screen, keyboard, and mouse. [INDICATOR] This indicates the work is designed to run on a desktop or laptop system and may require precise cursor control, large displays, or locally executed software. *genre_poetry*: [CORE] Poetry is a literary form that emphasizes rhythm, sound, and condensed language. [INDICATOR] This indicates the work focuses on expressive language, structure, or visual arrangement rather than linear storytelling.

3.4. Baseline: One-Hot Encoding

As a baseline representation, categorical values were encoded using one-hot vectors. When a work was associated with a single value for a given feature, the corresponding dimension was set to one and all others to zero. In the case of multi-valued features (for example, a work categorized under multiple techniques), a multi-hot representation was used in which several dimensions could simultaneously take the value one. All feature-level vectors were subsequently concatenated to form a single high-dimensional sparse vector representing each work.

As one-hot encoding ignores conceptual relationships between categories, it provides a useful control condition against which to evaluate the contribution of semantic enrichment.

4. Comparative Analysis

To assess the representational differences between one-hot and semantic encoding, we do not rely on supervised clustering evaluation measures. Such metrics presuppose the existence of a ground-truth partition against which clustering results can be evaluated. This assumption does not hold in our context. Instead, as our objective is to determine whether semantic encoding yields a more meaningful representation, we adopt an intrinsic evaluation approach based on the geometry of the resulting vector spaces, informed by domain knowledge.

First, we computed centroid vectors for works associated with the genre labels *poetry*, *narrative*, and *poetry_and_narrative*. We use genre as a primary analytical category because, as a broad and conceptually encompassing dimension, it is expected to occupy more distinct and structurally differentiated regions in the semantic space than more specific features such as hardware requirements or reading processes. Moreover, we expect *poetry* and *narrative* to be more distant from one another than *poetry_and_narrative* is from either of them. In this sense, genre offers a simple but meaningful relational structure against which the ability of the representation to capture graded semantic proximity can be assessed.

For each label, the centroid was obtained by averaging the representations of all works assigned to that category. Pairwise cosine distances between these centroids were then measured in both the one-hot and semantic spaces (see table 1). In the one-hot representation, the distances between the works assigned with these three labels are uniform and fail to reflect conceptual overlap. In contrast, preliminary results in the semantic space indicate that *poetry_and_narrative* occupies an intermediate position. This suggests that semantic information

Label A	Label B	OH	Sem
poetry	narrative	0.184	0.105
poetry	poetry_narrative	0.175	0.058
narrative	poetry_narrative	0.182	0.064

Table 1: Pairwise cosine distances between genre centroids. The matrix captures the geometric organization of genres within the embedding space, with smaller distances reflecting closer semantic alignment.

is encoded in the embedding space in a manner consistent with interpretive intuition.

Second, we used database-level centroids to determine whether semantic encoding reduces artificial separation caused by divergent terminologies. The working hypothesis was that one-hot encoding may exaggerate differences due to mismatched vocabularies, whereas semantic representations may align conceptually similar categories even when labels differ, therefore allowing for more graded similarities to emerge. In principle, in the case of these two specific databases, there should not be any significant semantic differences between them, since they cover the same literary forms, unless Brazilian artists have distinct practices to other Latin American artists.

Because one-hot and semantic encodings differ substantially in their geometric properties, raw centroid distances are not directly comparable across representations. One-hot encoding produces sparse, orthogonal vectors that tend to inflate distances, whereas semantic embeddings generate denser, continuous spaces in which distances are typically compressed. To avoid conflating representational scale with structural differentiation, we therefore rely on normalized measures of centrality and displacement.

First, we computed the ratio to global dispersion, i.e. the cosine distance between a database centroid and the global centroid of the embedding space, divided by the mean distance of all works to the global centroid. Because the measure is normalized by overall dispersion, it allows comparison across embedding spaces that differ in scale. Lower values indicate a more central position of the database relative to the overall space.

Second, we computed the relative offset, which compares the distance between a database centroid and the global centroid to the average internal dispersion of that database (i.e. the mean distance of works within the database to their own centroid). Lower values indicate that the database does not form a distinct cluster and spread across the whole space (see table 2).

Both one-hot and semantic encoding position the databases close to the global centroid, indicating that neither database forms a strongly displaced subregion of the representational space. However, semantic encoding slightly reduces both measures, which aligns with the intuition that it should smooth the artificial orthogonality of one-hot encoding.

DB	RGD_OH	RGD_S	RO_OH	RO_S
ALDAL	0.079	0.065	0.087	0.070
ALDB	0.122	0.097	0.130	0.104

Table 2: Normalized centrality of database centroids in one-hot and semantic embedding spaces. The ratio to global dispersion (RGD) quantifies relative centrality within each representational geometry, while the relative offset (RO) controls for internal database heterogeneity. Together, these metrics distinguish structural displacement from global scale effects introduced by different encoding strategies.

5. Discussion and Conclusions

The purpose of this study is to examine semantic encoding as a methodological intervention in digital humanities research. Humanities datasets are often small, curatorially constructed, and conceptually heterogeneous. In such contexts, representation is not merely a technical preprocessing step but an epistemological decision that shapes analytical outcomes.

One-hot encoding reflects a structuralist view of categorical metadata in which labels function as discrete and unrelated symbols. In doing so, it suppresses nuance, hybridization, and gradience. Semantic encoding, by contrast, embeds categorical labels within a continuous vector space informed by external linguistic knowledge. As shown by the genre experiment, this transformation enables composite categories to occupy intermediate positions and allows conceptually related labels to cluster together. It also smooths the artificial differences brought in by the use of different terminologies, as suggested by the database experiment.

By incorporating semantic embeddings derived from large language models, we introduce a representational layer that captures conceptual similarity beyond local statistical evidence. In doing so, we mitigate sparsity effects, facilitate cross-database comparison, and create a space in which interpretive proximity can be examined quantitatively.

Ultimately, semantic encoding should be understood as a representational strategy that complements, rather than replaces, symbolic metadata. It provides a way to bridge the gap between qualitative interpretive categories and quantitative analytical methods, allowing digital humanities researchers to explore conceptual structure without discarding nuance.

Future research should extend both methodology and the analysis enabled by semantic encoding. We followed the ARISE methodology regarding the attention-weighted pooling, as well as the choice of LLM and encoding model. We could investigate other options, as well as combine semantic encoding with insights from co-occurence patterns, as proposed in Yang et al. (2026). In addition, regarding the evaluation, while centroid-based measures offer an interpretable first approximation of structural organization, additional diagnostics could be implemented. For instance, neighborhood-based metrics, such as k-nearest-neighbor, could further assess whether semantic encoding enhances local conceptual coherence.

In the context of this specific use case, the semantically aware representations made possible by semantic encoding open several future research avenues, not only as an exploratory tool but as a methodological framework for investigating both the structure of the field of digital literature and the practices of database curation. On the one hand, such representations can be used to address research questions about the field itself. Diachronic analyses may track the evolution of practices over time, while demographic metadata (e.g., gender or country of origin of the authors) could be mapped onto the embedding space to explore patterns of diverging practices. On the other hand, they provide a means to critically assess database construction by identifying representational biases, such as whether certain databases disproportionately foreground particular types of works or authors while neglecting others.

Extending this methodology to the numerous digital literature databases which exists would make it possible to move toward a more comprehensive, bird's-eye view of the field. However, such comparisons would likely require methodological adaptations beyond those presented here, as differing feature sets and levels of granularity introduce additional challenges for alignment and semantic integration.

Finally, combining geometric analysis with qualitative inspection of representative works would help ground embedding-based findings in close reading, reinforcing the interpretive relevance of semantic encoding within digital humanities research.

6. References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases.

Manaíra Aires Athayde and REJANE ROCHA. 2022. Um arquivo para a literatura digital brasileira e algumas questões concretas. *RE-AUTO-META ARQUIVO*, page 84.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, and et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Constanza Botero Betancur. 2023. Gainza, c. y zúñiga, c. cartografía de la literatura digital latinoamericana. visualización, archivo y preservación de obras literarias digitales, 2018-2021. *publicaciones*, 4:e054.

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. Language models are few-shot learners. In *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 34, pages 1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zijin Feng, Luyang Lin, Lingzhi Wang, Hong Cheng, and Kam-Fai Wong. 2024. LLMEdgeRefine: Enhancing text clustering with LLM-based boundary point refinement. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18455–18462, Miami, Florida, USA. Association for Computational Linguistics.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. In *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, volume 35, pages 18932–18943.

Cheng Guo and Felix Berkhahn. 2016. Entity embeddings of categorical variables. *CoRR*, abs/1604.06737.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David A. Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pages 5549–5581. PMLR.

Karim Huesmann and Lars Linsen. 2025. Large language models for transforming categorical data to interpretable feature vectors. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):5754–5771.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Andrew Piper. 2018. *Enumerations: Data and Literary Study*. University of Chicago Press.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Zihua Yang, Xin Liao, Yiqun Zhang, and Yiu ming Cheung. 2026. Bridging the semantic gap for categorical data clustering via large language models.